Strategies for a Centralized Single Product Multiclass $M/G/1$ Make-to-Stock Queue

Hossein Abouee-Mehrizi  
Department of Management Sciences, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, haboueem@uwaterloo.ca

Barış Balcioglu  
Faculty of Engineering and Natural Sciences, Sabancı University, Orhanlı-Tuzla, 34956 Istanbul, Turkey, balcioglu@sabanciuniv.edu

Opher Baron  
Joseph L. Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada, opher.baron@rotman.utoronto.ca

Make-to-stock queues are typically investigated in the $M/M/1$ settings. For centralized single-item systems with backlogs, the multilevel rationing (MR) policy is established as optimal and the strict priority (SP) policy is a practical compromise, balancing cost and ease of implementation. However, the optimal policy is unknown when service time is general, i.e., for $M/G/1$ queues. Dynamic programming, the tool commonly used to investigate the MR policy in make-to-stock queues, is less practical when service time is general. In this paper we focus on customer composition: the proportion of customers of each class to the total number of customers in the queue. We do so because the number of customers in $M/G/1$ queues is invariant for any nonidling and nonanticipating policy. To characterize customer composition, we consider a series of two-priority $M/G/1$ queues where the first service time in each busy period is different from standard service times, i.e., this first service time is exceptional. We characterize the required exceptional first service times and the exact solution of such queues. From our results, we derive the optimal cost and control for the MR and SP policies for $M/G/1$ make-to-stock queues.

Subject classifications: make-to-stock; $M/G/1$ queue; priority classes; customer composition; multilevel rationing; strict priority.

Area of review: Manufacturing, Service, and Supply Chain Operations.

History: Received June 2009; revisions received April 2010, May 2010, February 2011, July 2011, December 2011; accepted February 2012.

1. Introduction

Market segmentation and customer differentiation are widely accepted ways to increase profitability. A common way to differentiate among customers is to provide different service levels for different customer classes. For example, in a make-to-stock system, service level is often measured by product availability on the shelf. In this case, the service level is directly influenced by allocation policies and inventory levels. An important research and managerial question is whether customer classes requesting the same product should be prioritized and if so how to prioritize them. In our examination of this question, we analyze inventory control strategies for a supplier using a centralized inventory to serve a single product to $n$ classes of customers. Assuming that class 1 has the highest priority and class $n$ has the lowest priority, we model the underlying production system as an $M/G/1$ queue.

Many policies are available to handle production and inventory control. Broadly speaking, however, inventory control policies can be characterized by whether customer types are prioritized, and whether allocation decisions are made when production starts or are postponed until production is completed. In this paper, we focus on centralized inventory control policies with postponement of the allocation decision. Note that because postponing allocation provides extra information, it may result in the same or a lower total cost as not postponing.

We assume that demand that is not immediately satisfied from stock is backlogged. Similar to earlier literature, we consider a first-come, first-served (FCFS) policy analyzed by Sanajian and Balcioglu (2009) along with the following two centralized inventory control policies that use a base-stock level control for their production decision.

**MR policy.** Under a multilevel rationing policy, there are nondecreasing threshold inventory levels $R_r$, $r = 1, \ldots, n+1$ with $R_1 = 0$ and $R_{n+1} = S$. If the inventory level, $I$, is between $R_r + 1$ and $R_{r+1}$, i.e., $R_r < I \leq R_{r+1}$, only demand requests of classes 1 to $r$ are satisfied on a FCFS basis. If the inventory level is between $R_r + 1$ and $R_{r+1}$, even if there are pending orders from classes $r+1$ to $n$, the completed product is placed in inventory. When there is no
stock, a finished product is allocated to the highest-priority customer backlogged (in a FCFS fashion within this class). When the inventory reaches $R_{n+1}$, the base-stock level, production stops.

**SP policy.** The strict priority policy is a special case of MR policy when $R_1 = R_2 = \cdots = R_n = 0$. That is, as long as there is stock in the centralized inventory, demand requests are satisfied on a FCFS basis. When there are backlogs, a finished product is allocated to the highest-priority customer among those with pending orders in the system.

Ha (1997a, b) was the first to discuss inventory rationing problems in a centralized make-to-stock system with different classes of customers. For exponentially distributed production times, Poisson arrivals, and lost sales, Ha (1997a) shows that the MR policy is optimal. Ha (1997b) extends this work to the backlog case with two classes of customers and shows that a stationary critical-level policy is optimal. de Véricourt et al. (2002) show that the MR policy is the optimal policy for the $M/M/1$ make-to-stock queues. de Véricourt et al. (2001) introduce the strict priority (SP) policy and compare the FCFS, SP, and MR policies for an $M/M/1$ queuing system, and demonstrate that the MR policy outperforms the other two. Ha (2000) considers an $M/E_k/1$ make-to-stock system with lost sales, where $E_k$ denotes $k$-stage Erlang service time and characterizes the optimal stock allocation policy. Gayon et al. (2009) propose a heuristic to approximate these levels for systems with Erlangian service times. Applications of rationing inventory have also been investigated when supply is ample; see Arslan et al. (2007) and references therein.

In this paper, we consider the SP and MR policies for a centralized single product multiclass $M/G/1$ system. Although the characterization of the optimal FCFS policy in this setting is known, we are the first to consider the MR and SP policies. We focus on cases where the product allocation is postponed to the end of production when it is allocated to one customer, possibly according to the customer priority. Note that this allocation does not change the total inventory level, but may reduce costs. We ignore additional information, such as the length of time since the start of production of the current item, something which might be both available and valuable in $M/G/1$ settings. For example, both Ha (2000) and Gayon et al. (2009) consider Erlangian service times and use information on production status. Although not using additional information might increase the costs of these policies relative to the optimal control policy, it keeps implementation simple and increases practicality.

Observe that in the MR system, the rate of change of the inventory level varies dynamically according to the rationing levels; this also changes customer composition, i.e., the proportion of each customer class out of the total number of customers in this queue. Note that because the total number of customers is invariant for every nonidling and nonanticipating policy (for a rigorous definition of such policies see, e.g., Bertsimas 2012), the various controls only change customer composition.

To express customer composition under MR and SP policies, we consider a series of multipriority class $M/G/1$ queues. In these queues, the first service time in each busy period is different from other service times, i.e., these queues have exceptional first service times in their busy periods. We show that with a careful choice of the exceptional first service times, their customer composition will be the same as the original $M/G/1$ system.

We obtain closed-form expressions for the optimal cost and base-stock level for an $M/G/1$ make-to-stock system under the SP policy. We also derive a computational approach to obtain the optimal cost and rationing levels for the MR policy for an $M/G/1$ system, i.e., with general service times. Previous work found these optimal controls using dynamic programming for exponential (or Erlang), service times, but when the service times are not exponential, dynamic programming is less practical. For example, Gayon et al. (2009) highlighted the difficulty finding the optimal controls in $M/E_k/1$ settings when the number of customer types is large. However, because the customer composition methodology employs a series of queues it allows the solution of systems with numerous customer types, as we demonstrate numerically in §3.4.2. We also show that the cost of the SP system is equivalent to the cost of a FCFS system with an appropriately defined backlog cost. Our theoretical and numerical results support the applicability of both the SP and MR policies for single product multiclass $M/G/1$ systems.

Our solution for the SP and MR policies relies on (i) the exact analysis of a multipriority $M/G/1$ queue with postponement and exceptional first service times in its busy periods, and (ii) characterizing the relevant exceptional first service times. Because the derivation of both is technical and intricate, we only present it in EC.1 (in §1 in the electronic companion). An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/oper.1120.1062. In §2, we present the multiclass $M/G/1$ system and the terminology used in the paper. In §3, we derive the optimal rationing levels, base-stock levels, and costs of the FCFS, SP, and MR policies. The proofs of the main results in Theorems 1 and 2 appear in §4 and the rest of the proofs appear in EC.2.

## 2. Modeling a Single Product Multiclass $M/G/1$

The single product multiclass $M/G/1$ system we consider has a supplier that produces a single product and caters to demand arising from $n$ distinguishable classes. We assume that the demand of each class $r$ (type $r$ demand) follows a Poisson process with a rate $\lambda_r$, $r = 1, 2, \ldots, n$. We use the terms type $r$ and class $r$ interchangeably. We model the general production times as i.i.d with a mean $1/\mu$ and a second moment $m_2$. Let $b(\cdot)$ and $b(\cdot)$ denote the probability density function and its Laplace transform (LT), respectively.
We assume that unsatisfied demand is backlogged. Thus, for stability, we require \( \rho := \lambda / \mu < 1 \), where \( \lambda = \sum_{r=1}^{n} \lambda_r \).

The backlog cost of class \( r \) is \( b_r \) per unit backlogged per unit time. Without loss of generality, we assume that \( b_1 > b_2 > \cdots > b_n \) (if two distinct classes have the same backlog cost, we aggregate them to a single class). Customers are prioritized according to their backlog costs, i.e., classes 1 to \( n \) from highest to lowest. The system incurs a holding cost of \( h \) per unit per unit time.

This model gives rise to a multiclass system where the server can work on one production order at a time. For this problem, we consider a centralized continuously reviewed inventory system. We use a production control according to a base-stock level, \( S \): thus, production stops, and the server becomes idle when the inventory level reaches \( S \).

We consider three different systems, corresponding to three different production control policies: the FCFS, SP, and MR systems. (From now on, we use these short terminologies, e.g., “SP system” rather than “multiclass single-item \( M/G/1 \) make-to-stock system with postponement of the product allocation to the end of production under an SP control policy.”)

Let \( I(t) \) denote the inventory level at time \( t \) in the system, and note that \( I(t) < 0 \) implies a backlog in the system. Let \( B_r(t) \) be the number of type \( r \) backlogs in the system.

In the FCFS and SP systems, if any class is backlogged at time \( t \), we have \( I(t) < 0 \); then \( I(t) < 0 \) implies a backlog of size \( |I(t)| \). However, in the MR system, we can have positive inventory on hand while some customer classes are backlogged; thus, \( I(t) > 0 \) and \( \sum_{r=1}^{n} B_r(t) > 0 \) is possible.

A standard method to express \( I(t) \) in a single-class production system with base-stock level control, when only \( I(t) < 0 \) implies a backlog, is to consider the shortfall process \( N(t) := S - I(t) \), e.g., Baron (2008) and references therein. Then, \( N(t) \) is identical to the number of orders in an \( M/G/1 \) queue facing (a) allocation, (b) demand, and (c) service processes that are identical to those faced by the original system. A shortfall \( N(t) < S \) implies that the inventory in the system has \( S - N(t) \) units; a shortfall \( N(t) > S \) implies a backlog of \( |S - N(t)| = N(t) - S \) units.

We use the shortfall queue to match the original FCFS and SP inventory systems to a queueing model.

We use a reasoning similar to the one that guides the use of the shortfall queue when analyzing the three systems mentioned. That is, we derive the cost of each system by analyzing a multiclass \( M/G/1 \) queue with the same allocation, demand, and service processes as in the original system.

An important observation with respect to the shortfall process, \( N(t) \), is that it is invariant under all nonidling and nonanticipating control policies. Because \( N(t) \) is invariant, we have

\[
N(t) = (S - (I(t))^+) + \sum_{i=1}^{n} B_i(t),
\]

where \( (x)^+ := \max(0, x) \).

Earlier we defined customer composition as the proportion of each customer class in the total number of customers in a queue. Given Equation (1), knowing the customer composition resulting from specific priorities and allocation rules in this queue is sufficient to represent the cost of this control for the relevant system. To express the relevant customer compositions in the SP and MR systems when they have a backlog, we construct multiclass single-item \( M/G/1 \) queues with postponement of allocation and exceptional first service times in busy periods. We name these queues “backlog queues” for simplicity. We will elaborate upon the ideas of customer composition and backlog queues in the next section.

3. The Costs and Optimization of the Three Policies

We use the backlog queues to derive the exact cost of the SP and MR systems in §3.2 and 3.3, respectively. For the sake of completeness (and better comparative analysis), in §3.1 we begin with the optimal control and corresponding cost for the FCFS system. We compare the performances of the three systems in §3.4.

The solution of multiclass \( M/G/1 \) queues with exceptional first service times and the derivation of the LT of the required exceptional service times are presented in EC.1.

3.1. The FCFS Policy

Recall that \( N(t) \) denotes the number of orders in the shortfall queue at time \( t \). Let \( P(i) := P(N = i) \) be the steady-state probability of having \( i \) orders in the shortfall queue.

Because all customers are treated the same, the average backlog cost per customer is \( b^f := \sum_{i=1}^{n} \lambda_i b_i / \lambda \). Therefore, for a given base-stock level \( S \), the average cost for the FCFS policy is

\[
C_f(S) := h \sum_{i=0}^{S} (S - i) P(i) + b^f \sum_{i=S+1}^{\infty} (i - S) P(i),
\]

and letting \( F(i) := \sum_{j=0}^{i} P(j) \), the optimal base-stock level, \( S^* \), that minimizes this cost is (see, e.g., Veatch and Wein 1996),

\[
S^* = \min \{ i : F(i) > b^f / (h + b^f) \}.
\]

Note that \( P(i) \) can be obtained in closed form using Equation (12) in Kerner (2008) after setting \( \lambda_i = \lambda \) as

\[
P(i) = (1 - \rho) \prod_{j=i+1}^{\infty} \frac{1 - \tilde{b}_j(\lambda)}{\tilde{b}(\lambda)}, \quad i = 1, \ldots,
\]

where \( \tilde{b}_j(\cdot) \) is the LT of the residual service time observed by an order arrival that sees \( j \) orders in the shortfall queue. This LT can be obtained recursively from Equation (4) in Kerner (2008):

\[
\tilde{b}_j(s) = \frac{\lambda}{s - \lambda} \left[ \tilde{b}(\lambda) \frac{1 - \tilde{b}_{j-1}(s)}{1 - \tilde{b}_{j-1}(\lambda)} - \tilde{b}(s) \right], \quad j \geq 1,
\]

where \( \tilde{b}_0(s) = \tilde{b}(s) \).
3.2. The SP Policy

We next express the cost of the SP system with a base-stock level $S$. Let $P(B_r = i)$ denote the steady-state probability of having $i$ backlogs from class $r$. The average cost for the SP system is

$$C_{SP}(S) = h \sum_{i=0}^{S} (S-i)P(i) + \sum_{i=0}^{n} b_r \sum_{i=0}^{\infty} iP(B_r = i)$$

$$= h \sum_{i=0}^{S} (S-i)P(i) + \sum_{i=1}^{n} b_i E[B_i],$$

(5)

where $E[B_i]$ is the expected number of backlogs of type $r$.

Observe that because the holding cost is independent of the classes, the shortfall queue is sufficient to express the holding cost in this system. When $N(t) > S$, the inventory in the system has $N(t) - S$ backlogs. But because the backlog costs differ among classes, the shortfall queue is insufficient to express these costs. We obtain $E[B_i]$ by constructing the SP backlog (SPB) queue. We then use $E[B_i]$ to characterize the optimal SP control policy and its corresponding cost.

3.2.1. The SP Backlog Queue. We construct the SPB queue to obtain the probabilistic description of the shortfall queue during periods with no inventory. To differentiate between queues, we use the terms customers in the SP system, orders in the shortfall queue, and job in the SPB queue.

We construct the SPB queue by specifying its (a) allocation, (b) arrival, and (c) service processes. As proved in Theorem 1, our construction ensures that the job composition in the SPB queue will match the customer composition in the SP system when there is no inventory in the system, i.e., when $N(t) \geq S$.

**Step (a).** At the end of each service completion, the SPB queue will remove the oldest job with the smallest $r$ index, making it a priority queue with the same priorities as the SP system when it has no inventory.

**Step (b).** The arrival process of jobs of type $r$ to the SPB queue will follow a Poisson process with rate $\lambda_r$, $r = 1, 2, \ldots, n$. Thus, the arrival processes for the SP system and the SPB queues are identical (in distribution).

**Step (c).** The first service time in each busy period of the SPB queue will be the equilibrium (steady-state) residual service time observed by an order arrival who finds orders in the shortfall queue upon its arrival. We let $b^{SPB}_0(\cdot)$ denote the LT of this equilibrium residual service time. When an exceptional first service time has ended, if there are other jobs in the SPB queue, all service times until the SPB queue clears all its jobs, follow a regular service distribution, with a LT $\tilde{b}(\cdot)$.

We set the service process to include the exceptional first service time in step (c) because every order arrival that sees $S$ orders in the shortfall queue creates a backlog. Thus, the service times of the first jobs in the busy periods of the SPB queue are identical in distribution to the residual service times of the customers in service in the SP system once a period with backlog starts.

To summarize, our construction in steps (a–c) indicates that the SPB queue is an $M/G/1$ priority queue with postponement and exceptional first service times in its busy periods. These exceptional first service times have a LT $\tilde{b}^{SPB}(\cdot)$ identical to the LT of the equilibrium residual service times observed by an arrival to the shortfall queue that sees $S$ orders in front of it. The LT of the other service times is of regular service times, $\tilde{b}(\cdot)$.

Let $P^{SPB}_r(i)$ denote the steady-state probability of having $i$ jobs of class $r$ in the SPB queue. We next state our first main result. The proof is given in §4.

**Theorem 1.** The steady-state probability of having $i$ backlogs from class $r$ in the SP system is

$$P(B_r = i) = [1 - F(S - 1)] P^{SPB}_r(i),$$

$$r = 1, 2, \ldots, n, \quad i = 1, \ldots, S$$

(6)

By Theorem 1 the probability of having $n$ type $r$ backlogs in the SP system is identical to the probability of having $n$ type $r$ jobs in the SPB queue given the system is out of stock. The latter depends of course on $b^{SPB}_0(\cdot)$. Although the theorem does not provide these probabilities, they are not required to express the cost function given in Equation (5), all we need is the expected number of type $r$ backlogs in the system. Given Theorem 1, given the system is out of stock, this expectation is identical to the expected number of type $r$ jobs in the SPB queue. Thus, we next characterize it.

The customer composition in the SPB queue is an essential building block in our analysis of the SP and MR systems and is given in Theorem 2. The proof of the theorem requires the derivations from EC.1 and is given in §4.

**Theorem 2 (Customer Composition).** The ratio of expected number of type $r$ customers, $E[N^{SPB}_r]$, to the expected number of total customers, $E[N^{SPB}]$, in the SPB queue is

$$E[N^{SPB}_r] = \rho \left( \frac{1}{\lambda_r^+} - \frac{1}{1 - \rho_r^+} \right),$$

$$E[N^{SPB}]$$

(7)

where $\lambda_r^+ := \sum_{i=1}^{r} \lambda_i$, and $\rho_r^+ := \lambda_r^+ / \mu$ for $r = 1, \ldots, n$.

Observe that, surprisingly, the ratio in Equation (7) is independent of $b^{SPB}_0(\cdot)$, and this ratio only depends on the first moments of the queue’s arrival and service processes.

3.2.2. Deriving the Optimal SP Policy. de Véricourt et al. (2001) show that the optimal cost of the SP system in the $M/M/1$ settings can be obtained by considering a FCFS single-class $M/M/1$ queue with a specific backlog cost. Theorem 3 uses Theorems 1 and 2 to extend this result to the $M/G/1$ system and show that the specific backlog cost only depends on the first moment of the (regular) service time.
Theorem 3 (Optimal SP Policy). The cost of the SP policy with base-stock level \( S \) is the same as that of a FCFS single-class \( M/G/1 \) queue with weighted backlog cost:

\[
b_{SP} = \sum_{i=0}^{S-1} \lambda_i (1 - \rho_i) b_i = \sum_{i=S}^{\infty} \lambda_i (1 - \rho_i^* ) (1 - \rho_{i-1}^* ).
\]

(8)

Thus, the cost of the SP policy can be written as

\[
C_{SP}(S) = h \sum_{i=0}^{S-1} (S - i)P(i) + b_{SP} \sum_{i=S}^{\infty} (i - S)P(i),
\]

(9)

and the optimal base-stock level \( S^{SP} \) that minimizes Equation (9) is

\[
S^{SP} = \min_i \{ F(i) > b_{SP} / (h + b_{SP}) \}.
\]

Observe that according to Theorem 3, finding the optimal base-stock level and cost of the SP system requires only the solution of a standard single-class FCFS \( M/G/1 \) queue. More specifically, we do not need to solve the SPB queue or characterize its exceptional first service times. Therefore, we find \( C_{SP}(S^{SP}) \) as we found \( C_{SP}(S^{SP}) \) by setting the backlog cost to \( b_{SP} \), as given in Equation (8), and we express \( S^{SP} \) and its corresponding cost using Equations (10) and (9), respectively.

3.3. The Multilevel Rationing Policy

Let \( C_{MR} := C(R_1 = 0, R_2, \ldots, R_{n+1} = S) \) be the long-run average cost of the MR system given rationing levels \( R_1, R_2, \ldots, R_{n+1} \). In this section, we derive the closed-form expression for this cost. The idea in developing this expression is similar to the one used for analyzing the SP policy. Specifically, we derive the customer composition within each relevant inventory range, \( I(t) \in (R_i, R_{i+1}] \) for \( i = 1, \ldots, n \) and \( I(t) \leq 0 \), by considering a properly defined backlog queue.

The proof of the following corollary relies on Theorem 2.

Corollary 1. We can assume without loss of generality that \( R_i > R_{i-1} \) for \( r = 2, \ldots, n + 1 \).

3.3.1. The MR Backlog Queues. Here we construct a series of backlog queues for each class \( r = 1, \ldots, n + 1 \). We denote class \( r \) backlog queue by \( BQ_r \). In Theorem 5 we show that the job composition in the backlog queues is identical to the relevant customer composition in the MR system.

We constructed the SPB queue by carefully constructing its (a) allocation, (b) arrival, and (c) service processes when \( I(t) \leq 0 \). We follow steps (a)–(c) next, formulating \( BQ_r \) for \( I(t) \leq R_r \), as a two-priority \( M/G/1 \) queue with postponement and exceptional first service times in its busy periods.

Step (a). We set \( BQ_r \) as a two-priority queue in which priority is given to the jobs of classes \( 1, \ldots, r - 1 \) over jobs of class \( r \).

The intuition behind step (a) is that once the inventory hits a rationing level and the customer composition changes, only the priority of a single class of customers changes; all other classes are treated as before. For example, once the inventory falls below \( R_r + 1 \), classes 1, \ldots, \( n - 1 \) remain high priority, receiving items from inventory upon arrival; and only the priority of class \( n \) customers changes from high to low. Therefore, \( BQ_n \) is a two-priority queue in which jobs of types 1, \ldots, \( n - 1 \) are high priority, and jobs of type \( n \) are low priority.

Step (b). We set the arrival processes of all job types to be Poisson, and let the high- and low-priority jobs arrival rates for \( BQ_r \) be \( \lambda_{r-1}^+ := \sum_{j=r}^{n} \lambda_j \) and \( \lambda_r \), respectively. This queue ignores customers of classes \( r + 1, \ldots, n \).

We set the arrival rates of the low- and high-priority jobs in \( BQ_r \) as defined in step (b) because of the following:

Observation 1. For any class \( r = 2, \ldots, n \), once the inventory level in the original system decreases to \( R_r \), type \( r \) customers become low-priority until the inventory climbs to \( R_r + 1 \) again. During these periods the inventory level might downcross \( R_j \) for other classes \( j < r \), making them low-priority customers and backlogging their demand. It is possible that all stock will be depleted and all demand backlogged. However, before the inventory climbs to level \( R_r \), the system first clears the backlogs of classes \( j < r \). In other words, from the point of view of class \( r \), classes 1 to \( r - 1 \) remain a single class of high-priority customers as long as \( I(t) \leq R_r \). Similarly, as long as \( I(t) \leq R_r \), classes \( j > r \) are low-priority and, therefore, their arrivals do not affect the system times experienced by classes \( j \leq r \).

Observation 1 implies that any change of class \( r \) backlog in the MR system corresponds to a change of the low-priority job in \( BQ_r \) and to a change in the high-priority job in \( BQ_j \) for \( j > r \). However, this change of the class \( r \) backlog does not affect \( BQ_j \) for \( j < r \). Thus, we ignore class \( r \) when considering \( BQ_j \) for \( j < r \), i.e., the backlog queues of higher priority classes.

Step (c). We set the service process of the \( BQ_r \) to have exceptional first service times in busy periods and regular service times with LT of \( \tilde{b}_r(\cdot) \) otherwise. We set the LT of the exceptional service times to be the LT of the residual service times observed by a high-priority arrival at \( BQ_{r+1} \) that sees \( R_{r+1} - R_r \) jobs in the queue. We let \( \Delta_r := R_{r+1} - R_r \) for \( r = 1, \ldots, n \) and denote this LT by \( \tilde{b}_r^E(\cdot) \). (With this notation, \( h_r^E(\cdot) \) is identical to \( b_r^{SP}(\cdot) \), the LT of the exceptional first service times in the SPB queue.)
system to \( R_r \). Consider high-priority job arrivals in \( BQ_{r+1} \) that see \( R_{r+1} - R_r \) high-priority jobs. Every such arrival corresponds to a customer that decreases the inventory in the system to \( R_r - 1 \) or creates a class \( r \) backlog. With our construction, every such high-priority arrival corresponds to jobs that start the busy period in \( BQ \). Therefore, we set the first service times in busy periods in \( BQ \), as the equilibrium residual service times observed by high-priority arrivals that see \( R_{r+1} - R_r \) high-priority jobs in \( BQ_{r+1} \). This choice makes the service time of the first jobs in busy periods of \( BQ \) identical, in distribution, to the required residual service times. As Theorem 5 below states, this construction, together with steps (a) and (b), results in that the job composition in \( BQ \) is identical to the relevant (classes 1, \ldots, \( r \)) customer composition in the MR system when \( f(t) \leq R_r \).

To summarize, for \( r = 2, \ldots, n \), \( BQ_r \) is a two-priority \( M/G/1 \) queue with high- and low-priority customer arrival rates \( \lambda^+_r = \sum_{i=1}^{r-1} \lambda_i \) and \( \lambda_r \), respectively, and exceptional first service times in its busy periods. The LT of the first service times is \( \tilde{b}^r_n(\cdot) \) and the LT of regular service times is \( \tilde{b}(\cdot) \).

For completeness, we think of the shortfall queue of the MR system as the \( n + 1 \) backlog queue, \( BQ_{n+1} \). We let \( \lambda_{n+1} := 0 \) and set the first exceptional service times to be regular service times with a LT \( \tilde{b}^{n+1}_0(\cdot) = \tilde{b}(\cdot) \). This implies that all jobs in \( BQ_{n+1} \) form a single high-priority class.

Note that we can calculate the backlog of class 1 customers from \( BQ_2 \) (this is \( (i - R_i)^+ \), where \( i \) is the number of high-priority customers in \( BQ_2 \)). However, as shown in Theorem 2, finding the expected number of customers in a backlog queue can be done in closed form. Thus, to reduce the computational burden, we use \( BQ_1 \). The exceptional first service times for this queue are the residual service times seen by high-priority arrivals at \( BQ_2 \) that see \( \Delta_i \), high-priority jobs in the queue, i.e., its LT is \( \tilde{b}^2_{\Delta_i}(\cdot) \).

Let \( \rho_b \) be the server utilization in \( BQ_r \), and \( 1/\mu_1 \) and \( m^1_1 \) be the first and second moments of the exceptional first service times in \( BQ_r \). Both \( 1/\mu_1 \) and \( m^1_1 \) can be obtained using \( \tilde{b}^1_{\Delta_1}(\cdot) \) that can be derived using Theorem 2.1.2 presented in EC.1.2. (For notational convenience and because the context is clear, we omit the superscript \( r \) from \( \rho_b, \mu_1 \), and \( m^1_1 \) in \( BQ_r \).) Because Poisson arrivals see time averages, the mean of service time is \( 1/\mu_1 \) with probability \( \rho_b \) and \( 1/\mu_1 \) with probability \( 1 - \rho_b \), thus

\[
\rho_b = \frac{\lambda^+_1(1 - \rho_b)}{\mu_1} + \frac{\lambda_1 \rho_b}{\mu} = \frac{\lambda^+_1 \mu}{\mu_1 \mu + \lambda^+_1 (\mu - \mu_1)}. \tag{11}
\]

Let \( E[N^{BQ}_1] \) and \( E[N^{BQ}_r] \) denote the expectation of the number of all (total) and low-priority jobs in \( BQ_r \), respectively. Also, for \( r = 1, \ldots, n + 1 \), let \( P^1_{BQ}(i) \) and \( P^r_{BQ}(i) \) denote the steady-state probability of having \( i \) high- and low-priority jobs in \( BQ_r \), respectively. Next we derive closed-form expressions for \( E[N^{BQ}_1] \) and \( E[N^{BQ}_r] \) and generalize Equation (2), which is given for the distribution of the total number of orders in a FCFS \( M/G/1 \) queue, to the distribution of the number of high-priority jobs in \( BQ_r \). We define \( \Pi^k_{BQ}(\cdot) := 1 \) for \( k > 1 \).

**Theorem 4.** Consider \( BQ_r \).

1. The expected number of type \( r \) jobs in \( BQ_r \), is

\[
E[N^{BQ}_r] = \sum_{j=0}^{\infty} j P^r_{BQ}(j) = E[N^{BQ}_r] \frac{\lambda^+_r}{\lambda_r} \frac{1 - \rho^+_r}{(1 - \rho^+_r)(1 - \rho^-_r)}, \tag{12}
\]

where \( \rho^+_r = \lambda^+_r/\mu \) for \( r = 1, \ldots, n \), \( \rho^+_0 := 0 \), \( \lambda^+_r = \sum_{i=1}^{r} \lambda_i \) as before, and

\[
E[N^{BQ}_0] = (1 - \rho_b) \lambda^+_1 \frac{(\lambda^+_1)^2 \mu_1 + (1 - \rho)(\lambda^+_1 m^1_1 + 2/\mu_1)}{2(1 - \rho^2)}, \tag{13}
\]

2. The probability of having \( i \) high-priority jobs in \( BQ_r \), is

\[
P^r_{BQ}(i) = \frac{(\rho_b - \lambda, E[A])^{-1} - \tilde{b}^{r-1}_i(\lambda^+_{r-1})}{\lambda^+_r} \prod_{j=0}^{r-1} \frac{\tilde{b}^{r-1}_i(\lambda^+_{r-1})}{b(\lambda^+_{r-1})}, \tag{14}
\]

where \( \rho_b \) is given in Equation (11), \( E[A] \) is given in Lemma EC.2, and \( \tilde{b}^{r-1}_i(\cdot) \) are given in Theorem EC.2.

The proof of Theorem 4 relies on Theorem 2 and uses a similar derivation to that in Kerner (2008). Note that \( P^r_{BQ}(i) \) is a function of \( \tilde{b}^{r-1}_i(\cdot) \) that can be obtained recursively using Algorithm 1 given in EC.1.3 starting with \( \tilde{b}^{n+1}_0(\cdot) = \tilde{b}(\cdot) \).

Finally, the system’s inventory and backlog probabilities can be obtained from \( BQ_r \) with \( j = 2, \ldots, n + 1 \), and \( j = 1, \ldots, n \), respectively, as given in Theorem 5. Although the proof of Theorem 5 is similar to the proof of Theorem 1 for the SP system, it requires more work. The proof ties \( BQ_r \) to \( BQ_{r+1} \) using induction, and then ties \( BQ_r \) to the MR system. Table 1 summarizes the relations between these queues and the MR system.

Let \( F^r_{BQ}(i) := \sum_{j=0}^{i-1} P^r_{BQ}(j) \) and \( \tilde{F}^r_{BQ}(i) = 1 - F^r_{BQ}(i) \).

**Theorem 5.** (i) The steady-state probability of having \( i \) backlogs from class \( r \) in the MR system is

\[
P(B_r = i) = \prod_{j=r+1}^{n+1} F^r_{BQ}(\Delta_{j-1} - 1) P^r_{BQ}(i), \tag{15}
\]

where \( \Delta_i \) is the total number of orders in a FCFS \( M/G/1 \) queue, to the distribution of the number of high-priority jobs in \( BQ_r \). We define \( \Pi^k_{BQ}(\cdot) := 1 \) for \( k > 1 \).

(ii) The steady-state probability of having \( R_r - i \) inventory units in the MR system is

\[
P(I = R_r - i) = \prod_{j=r+1}^{n+1} F^r_{BQ}(\Delta_{j-1} - 1) P^r_{BQ}(i), \tag{16}
\]

where \( \Delta_i \) is the total number of orders in a FCFS \( M/G/1 \) queue, to the distribution of the number of high-priority jobs in \( BQ_r \). We define \( \Pi^k_{BQ}(\cdot) := 1 \) for \( k > 1 \).
3.3.2. The Cost of the MR Policy. Here we express $C_{MR}$ in closed form using the backlog queues. Combining Theorems 4 and 5, the total cost of the MR system is (no further proof is provided) the following:

**Theorem 6.** The long-run average cost of the MR policy is

$$
C_{MR} = h \sum_{r=0}^{n+1} \sum_{j=0}^{n+1} \prod_{i=r+1}^{j} F_{BQ}^{BQ}(\Delta_{j-i} - 1) \left( R_{r-i} - 1 \right) F_{BQ}^{BQ}(i) + \sum_{r=1}^{n} b_{r} \left[ \prod_{i=r}^{n+1} F_{BQ}^{BQ}(\Delta_{j-i} - 1) E[N_{r}^{BQ}] \right].
$$

Equation (17)

We remind that $E[N_{r}^{BQ}]$ and $F_{BQ}^{BQ}(i)$ are given in closed form in Theorem 4. To calculate Equations (13) and (14), we obtain the LTs of the exceptional first service times in $BQ_r$ by recursively using Theorem EC.2. Although the cost is in closed form, the expression is quite cumbersome because it uses the LT of different equilibrium residual service times.

3.3.3. Searching for the Optimal MR Policy. For a given set of rationing levels $R_1, \ldots, R_{n+1}$, if $R_i = R_{i+1} = \ldots = R_r$, we first aggregate customers of classes $i, \ldots, j$ as a single class and normalize their backlog costs using Theorem 2. Then, we calculate the cost of the MR system using Theorem 6. We start with $BQ_{n+1}$ that is a FCFS $M/G/1$ queue with an arrival rate $\lambda = \sum_{i=1}^{n} \lambda_i$ and obtain the probabilities $P_{BQ}(i)$ using Equation (4). To obtain the probabilities $P_{BQ}^{BQ}(i)$ for $BQ_2, \ldots, BQ_{n+1}$, we use Equation (14) that requires $B_{r+1}^y(i)$. We calculate these LTs using Theorem EC.2. Finally, we obtain $E[N_{r}^{BQ}]$ using Theorem 4 without calculating $P_{BQ}(i)$. With the exact cost $C_{MR}$ calculated using this procedure for given rationing levels, we can search over different vectors of $(R_1, R_2, \ldots, R_{n+1})$ to find the optimal rationing levels and the corresponding cost.

3.4. Comparison of the Three Policies

To compare the MR, SP, and FCFS $M/G/1$ systems, as before, we let $C_{SP}(S^{SP})$, $C_{SP}(S^{SP})$, and $C_{MR}$ denote the optimal cost of the FCFS, SP, and MR systems, respectively.

3.4.1. Theoretical Comparison. Note that the SP control is a special case of the MR control and that the customer composition in the SP system leads to lower backlog costs than in the FCFS system while maintaining the same holding cost. Observation 2 summarizes this and provides theoretical support for the use of the MR and SP policies rather than the FCFS policy in $M/G/1$ make-to-stock queues. The observation is given without a more detailed proof.

**Observation 2.** We have $C_{SP}(S^{SP}) \leq C_{MR}(S^{SP}) \leq C_{FCFS}(S^{SP})$.

3.4.2. Numerical Comparison. Our methodology can be used to find the optimal control and cost for 2, 5, and 10 customer classes. Because $M/M/1$ make-to-stock systems have been investigated (de Véricourt et al. 2001), we consider two service times with a squared coefficient of variation (variance to squared mean ratio) $cv^2 \neq 1$: (i) deterministic, with a mean of one and $cv^2 = 0$, and (ii) the two-stage mixed generalized Erlang (MGE2) distribution with $cv^2 = 2$, $MGE2 (\mu_1 = 1.05523, \mu_2 = 0.99477, a_1 = 0.99504)$ (Altuk 1997, pp. 42–43), a mean of one and density

$$
f(y) = \frac{(1 - a_1)}{\mu_1} \frac{(1 - a_2)}{\mu_2} \mu_1 e^{-\mu_1 y} + \frac{a_1}{\mu_1} \frac{a_2}{\mu_2} \mu_2 e^{-\mu_2 y}.
$$

We vary $\rho = 0.8, 0.9$ while maintaining the arrival rates equal $\lambda = \rho / n$, letting $b_r = n - r + 1$, $r = 1, \ldots, n$ (i.e., $b_n = 1$) and $h = 0.1$. This gives a total of 24 tests. For each test we calculated the ratios

$$
\Delta SP := \frac{C_{SP}(S^{SP}) - C_{MR}^*}{C_{MR}} \times 100,
$$

$$
\Delta F := \frac{C_{FCFS}(S^{SP}) - C_{MR}^*}{C_{MR}} \times 100.
$$

**Table 1.** Relations between backlog queues and the MR system.

<table>
<thead>
<tr>
<th>Queue is relevant</th>
<th>once the total number of high-priority jobs in the $(r + 1)$st backlog queue increases to $\Delta_r$.</th>
<th>The residual service time of a high-priority job that finds $C_{SP}(S^{SP})$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first service time in a busy period corresponds to</td>
<td>the residual service time of a high-priority job in this queue upon arrival.</td>
<td>the residual service time of a customer arrival of classes $1, \ldots, r$.</td>
</tr>
<tr>
<td>The busy period starts (and the idle period ends) with a job arrival that corresponds to</td>
<td>a high-priority job arrival to this queue that sees $\Delta_r$ high-priority jobs upon arrival.</td>
<td>a customer arrival that decreases $I(t)$ to $R_r - 1$ when $B_r(t) = 0$ or increases $B_r(t)$ to 1 when $I(t) = R_r$.</td>
</tr>
<tr>
<td>The busy period ends (and the idle period starts), corresponds to</td>
<td>a service completion that reduces the total number of high-priority jobs in this queue to $\Delta_r$.</td>
<td>when the inventory increases to $R_r$, while $B_r(t) = 0$ or when the class $r$ backlog decreases to 0 (this can only happen while $I(t) = R_r$).</td>
</tr>
</tbody>
</table>

Low-priority customers: | The lowest high-priority jobs in this queue. | Customers of class $r$. |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High-priority customers:</td>
<td>All but the lowest high-priority jobs in this queue, i.e., jobs of classes $1, \ldots, r - 1$.</td>
<td>Customers of classes 1 to $r - 1$.</td>
</tr>
</tbody>
</table>

[55x104]orem EC.2. Finally, we obtain $E_h^{BQ}$ without calculating $\tilde{\eta}$ that requires $\bar{r}$ and $\eta$. Cumbersome because it uses the LT of different equilibrium in Equation (17) is a closed-form expression, it is quite burdensome because it uses the LT of different equilibrium residual service times.
Table 2. \(\Delta SP\) and \(\Delta F\) for multiple classes of customers.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(cv^2)</th>
<th>(\Delta SP)</th>
<th>(\Delta F)</th>
<th>(\Delta SP)</th>
<th>(\Delta F)</th>
<th>(\Delta SP)</th>
<th>(\Delta F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.00</td>
<td>9.73</td>
<td>1.80</td>
<td>22.45</td>
<td>4.11</td>
<td>18.38</td>
</tr>
<tr>
<td>5</td>
<td>0.9</td>
<td>0.00</td>
<td>12.00</td>
<td>2.36</td>
<td>26.48</td>
<td>5.62</td>
<td>30.14</td>
</tr>
<tr>
<td>10</td>
<td>2.0</td>
<td>0.00</td>
<td>11.93</td>
<td>1.72</td>
<td>21.95</td>
<td>3.11</td>
<td>20.39</td>
</tr>
<tr>
<td>10</td>
<td>0.9</td>
<td>0.00</td>
<td>13.00</td>
<td>1.98</td>
<td>27.85</td>
<td>4.02</td>
<td>29.87</td>
</tr>
</tbody>
</table>

Table 2 presents the results of these numerical experiments and shows that using the MR and SP policies can significantly reduce costs, compared to the optimal FCFS policy.

4. Proofs of the Main Results

In this section we provide the proofs of our two main results. In Theorem 1, we show that the distribution of the number of customers in an \(M/G/1\) queue with priorities that depend on the number of customers in the system can be deduced by investigating a multipriority \(M/G/1\) queue with an exceptional service time. In Theorem 2, we characterize the cost composition in such queues.

**Proof of Theorem 1.** We first prove that the steady-state distribution of number of jobs in the SPB queue is identical to the steady-state distribution of number of backlogs in the system given that the system is out of stock:

\[
P(S + i) = [1 - F(S - 1)]P^{SP}(i), \quad i = 0, 1, \ldots, \tag{19}
\]

where \(P^{SP}(i)\) denotes the steady-state probability of having \(i\) jobs in the SPB queue. We then establish that the job composition in the SPB queue is identical to the customers backlog composition in the SP system given that the system is out of stock.

Equation (19) states that \(P^{SP}(i)\), is identical to the steady-state probability of having \((S + i)\) orders in the shortfall queue given that \(N(t) \geq S\).

Using Equation (4), the steady-state probability of having \((S + i)\) orders in the shortfall queue is

\[
P(S + i) = P(0) \prod_{j=0}^{S+i-1} \frac{1 - \tilde{b}_j(\lambda)}{\tilde{b}(\lambda)} = P(0) \prod_{j=0}^{S+i-1} \frac{1 - \tilde{b}_{S+j}(\lambda)}{\tilde{b}(\lambda)}, \quad i = 0, 1, \ldots, \tag{20}
\]

We next obtain the steady-state probability of having \(i\) jobs in the SPB queue. The derivation is similar to the one for the \(M/G/1\) queue in Kerner (2008). We define \(q_i(\eta, \eta)\) as the probability that there are \(i\) jobs in the SPB queue, and remaining service time is \(\eta\) at time \(t\). Therefore, we have

\[
q_{i+d}(1, \eta) = q_i(1, \eta + dt)(1 - \lambda dt) + q_i(2, 0)g(\eta) dt + q_i(0, 0)\tilde{b}_i(\eta) dt, \quad i = 1, \tag{21}
\]

\[
q_{i+d}(i, \eta) = q_i(i, \eta + dt)(1 - \lambda dt) + q_i(i - 1, \eta + dt)\lambda dt + q_i(i + 1, 0)g(\eta) dt, \quad i > 1, \tag{22}
\]

where \(\tilde{b}_i(\eta)\) is the density of the first exceptional service times in the SPB queue. Then, similar to the proof of Lemma 3.1.3.1 in Kerner (2008) we have,

\[
P^{SP}(i) = P^{SP}(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{b}_j(\lambda)}{\tilde{b}(\lambda)}, \tag{23}
\]

where \(\tilde{b}_i(\eta)\) is the LT of the equilibrium residual service times observed by arrivals who find \(i\) jobs in the SPB queue.

Setting \(\lambda_1 = 0\), \(\lambda_2 = \lambda\), and \(\tilde{b}_i(\eta) = \tilde{b}_0(\eta) = \tilde{b}_f(\eta)\) (where the last equality follows by our construction in step (c) in Theorem 2.2 we get \(\tilde{b}_i(\eta) = \tilde{b}_{S+i}(\eta)\) for \(i = 1, 2, \ldots\) Therefore,

\[
P^{SP}(i) = P^{SP}(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{b}_{S+j}(\lambda)}{\tilde{b}(\lambda)}, \quad i = 0, 1, 2, \ldots \tag{24}
\]

We next show that Equation (19) holds for \(i = 0\). Let \(1/\mu_1\) denote the expected remaining service time of an order in service in the shortfall queue observed by an arrival who finds \(S\) orders in the shortfall queue (that is \(-d\tilde{b}_s(s)/ds|_{s=0} = 1/\mu_1\)). Sigman and Yechiali (2007, Equation (1)) show that

\[
\frac{1}{\mu_1} = \frac{1 - \rho}{\lambda P(S)} (1 - F(S)).
\]

So that,

\[
P(S) = \frac{1 - \rho}{\lambda/\mu_1 + 1 - \rho} (1 - F(S - 1)). \tag{25}
\]

Also as in Equation (11) the utilization of the SPB queue, \(\rho_b\), is

\[
\rho_b = \frac{\lambda \mu}{\mu_1 + \lambda (\mu - \mu_1)} = 1 - \frac{1 - \rho}{\lambda / \mu_1 + 1 - \rho}. \tag{26}
\]

Comparing Equations (25) and (26) we get

\[
P(S) = (1 - F(S - 1))(1 - \rho_b) = (1 - F(S - 1))P^{SP}(0). \tag{27}
\]

Therefore, Equation (19) holds for \(i = 0\). Substituting Equation (27) on the right-hand side of Equation (20) together with Equation (24) establishes Equation (19) for \(i \geq 1\).

We next establish that the job composition in the SPB queue is identical to the customers backlog composition in the SP system. Intuitively, considering Equation (19), we observe that given step (a) of the construction of the SPB queue, the job is allocated in the SPB queue in the same way as it is allocated in the SP system while \(N(t) \geq S\). Furthermore, given step (b) in the construction of the SPB queue, the job arrival process of type \(r\) in the SPB queue has the same distribution as the customer arrival process of type \(r\) in the SP system. Both observations together
with Equation (19) imply that the job composition in the SPB queue is identical to the customer composition in the SP system. This implication establishes Equation (6).

More formally, consider the continuous time Markov chain (MC) that represents the jobs’ distribution in a multiclass M/G/1 queue with exceptional first service times with a density of $b_0(\cdot)$. Let $\bar{N} = (L_1, \ldots, L_n)$ denote the vector of the number of jobs of classes 1, $\ldots$, $n$, $\bar{N}_k = \arg \min \{ L_r > 0 \}$, and $\bar{k} = \{ r : L_r > 0 \}$ is the set of classes with jobs waiting in the system. Then, similar to Equations (21) and (22) this MC is given by

$$h_{i+dt}(\bar{N}, \eta) = h_i(\bar{N}, \eta + dt)(1 - \lambda \, dt) + \sum_{k=1}^{\infty} \tilde{h}_i(\bar{N} + \bar{e}_k, 0) b(\eta) \, dt, \quad \sum_{j=1}^{\infty} L_j = 0,$$  
(28)

$$\tilde{h}_{i+dt}(\bar{N}, \eta) = \tilde{h}_i(\bar{N}, \eta + dt)(1 - \lambda \, dt) + \sum_{k=1}^{\infty} \tilde{h}_i(\bar{N} + \bar{e}_k, 0) b(\eta) \, dt$$  
$$+ \tilde{h}_i(\bar{N} - \bar{e}_k, \eta + dt) \lambda_i b_0(\eta) \, dt, \quad \sum_{j=1}^{\infty} L_j = 1,$$  
(29)

$$\bar{h}_{i+dt}(\bar{N}, \eta) = \bar{h}_i(\bar{N}, \eta + dt)(1 - \lambda \, dt) + \sum_{k=1}^{\infty} \bar{h}_i(\bar{N} + \bar{e}_k, 0) b(\eta) \, dt$$  
$$+ \sum_{k \in \bar{k}} \bar{h}_i(\bar{N} - \bar{e}_k, \eta + dt) \lambda_i b_0(\eta) \, dt, \quad \sum_{j=1}^{\infty} L_j > 1,$$  
(30)

where $h_i(\bar{N}, \eta)$ is the probability that there are $L_r$ jobs of class $r$ in the system, and remaining production time is $\eta$ at time $t$.

Next consider the continuous time Markov chain that represents the backlogs’ distribution in the SP system during the periods that the system is out of stock. Let $N = (B_1, \ldots, B_n)$ denote the vector of the backlogs of classes 1, $\ldots$, $n$, $e_r$ denote the $r$th unit vector, $\bar{N}_r = \arg \min \{ B_r > 0 \}$ and $\bar{k} = \{ r : B_r > 0 \}$ is the set of backlogged classes. Then, similar to Equations (21) and (22) this MC is given by

$$h_{i+dt}(N, \eta) = h_i(N, \eta + dt)(1 - \lambda \, dt) + \sum_{k=1}^{\infty} h_i(N + e_k, 0) b(\eta) \, dt, \quad \sum_{j=1}^{\infty} B_j = 0,$$  
(31)

$$\bar{h}_{i+dt}(N, \eta) = \bar{h}_i(N, \eta + dt)(1 - \lambda \, dt) + \sum_{k=1}^{\infty} \bar{h}_i(N + e_k, 0) b(\eta) \, dt$$  
$$+ \bar{h}_i(N - e_k, \eta + dt) \lambda_i b_0(\eta) \, dt, \quad \sum_{j=1}^{\infty} B_j = 1,$$  
(32)

where $h_i(N, \eta)$ is the probability that there are $B_r$ backlogs of class $r$ in the SP system given it is out of stock, and remaining production time is $\eta$ at time $t$. Note that in this MC, $b_0^{\text{SPB}}(\cdot)$ is independent of class $r$ backlogs because any arrival to the SP system that finds inventory level equals zero creates the first backlog and starts the backlog period.

Comparing Equations (31)–(33) with Equations (28)–(30), respectively, we observe that the MC representing the backlogs in the SP system given it is out of stock is identical to the MC that represents the number of jobs in an M/G/1 queue with exceptional first service times if $b_0(\eta) = b_0^{\text{SPB}}(\eta)$. Therefore, because the density of the first exceptional service times in the SPB queue is defined as $b_0^{\text{SPB}}(\eta)$, we observe that the MC representing the backlogs in the SP system given it is out of stock is identical to the MC that represents the number of items in the SPB queue, and consequently the distribution of backlogs of class $r$ given the SP system is out of stock is identical to the distribution of jobs of class $r$ in the SPB queue. Note that this discussion essentially establishes Equation (19) as well. The derivation of Equation (19) is given above as it provides the closed-form expression for these probabilities. □

We next give the proof of Theorem 2.

**Proof of Theorem 2.** Consider customers of classes $1, \ldots, r$ as high priority with an arrival rate $\lambda_r^+$. Let $E[N_r^+]$ and $E[N_r^-]$ denote the expected number of customers of classes 1, $\ldots$, $r$ and $r+1, \ldots, n$ in the SPB queue, respectively. We call high- and low-priority classes $r^+$ and $r^-$, respectively.

Using Little’s law, we have $E[N_r^{\text{SPB}}] = -\lambda \tilde{w}(s)|_{s=0}$, and $E[N_r^-] = -\lambda_r^- \tilde{w}_r^-(s)|_{s=0}$, where $\tilde{w}(s) = \tilde{w}(s + \lambda^+ s \sum \theta^+ r|1 - \theta_r^+(s))$ and note that in this case, $\lambda_r^+ = 0$ and $\lambda_r^- = \lambda$. Therefore, we have a single-class M/G/1 queue with exceptional first service times. Using Corollary EC.1 we get $\tilde{w}_r(s) = \tilde{w}_r(s)$ as given in Equation (EC.12). Because $E[N_r^{\text{SPB}}] = E[N_r^+] + E[N_r^-]$, we have

$$E[N_r^-] / E[N_r^{\text{SPB}}] = 1 - \frac{E[N_r^-]}{E[N_r^{\text{SPB}}]} = 1 - \frac{-\lambda_r^- \tilde{w}_r^-(s)|_{s=0}}{-\lambda \tilde{w}(s)|_{s=0}} = 1 - \frac{\lambda_r^- \tilde{w}_r^-(s + \lambda^+ s \sum \theta^+ r|1 - \theta_r^+(s)) (1 - \lambda_r^- \tilde{w}_r^-(s)|_{s=0})}{\lambda \tilde{w}(s)|_{s=0}}.$$


where as in Equation (EC.11) \( \theta^+_r(s) = \tilde{b}(s + \lambda^+_r(1 - \theta^+_r(s))) \). Because \( \theta^+_r(0) = 1 \), \( \tilde{w}(0) \) cancels out and then because \( \tilde{b}(s)|_{s=0} = 1/\mu \), we have

\[
\frac{E[N^+_r]}{E[N^+_{SPB}]} = 1 - \frac{\lambda^+_r}{\lambda}(1 - \lambda^+_r \theta^+_r(s)|_{s=0})
\]

\[
= 1 - \frac{\lambda^+_r}{\lambda} \left( 1 - \frac{\lambda^+_r \tilde{b}(s)|_{s=0}}{1 + \lambda^+_r \tilde{b}(s)|_{s=0}} \right)
\]

\[
= \frac{\lambda^+_r (1 - \rho)}{\lambda(1 - \rho^+_r)},
\]

(34)

where \( \rho^+_r = \lambda^+_r/\mu \) and \( \rho = \lambda/\mu \).

Now consider a second system with two classes of customers where the arrival rates of high- and low-priority customers are \( \lambda^+_{r-1} \) and \( \lambda^-_{r-1} \), respectively. The expected number of high-priority customers in this system is \( E[N^+_{r-1}] \). The expected number of customers of class \( r \) in the multipriority class can be expressed as

\[
E[N^+_{r}] = E[N^+_{r-1}] - E[N^+_{r-1}],
\]

Therefore,

\[
\frac{E[N^+_{r}]}{E[N^+_{SPB}]} = \frac{E[N^+_{r-1}] - E[N^+_{r-1}]}{E[N^+_{SPB}]}.
\]

Applying Equation (34) to the \( r^+ \) and \( (r-1)^+ \) customers and substituting it into Equation (35) and letting \( \rho_r = \lambda_r/\mu \), we have

\[
\frac{E[N^+_{SPB}]}{E[N^+_{SPB}]} = \frac{\lambda^+_r (1 - \rho)}{\lambda(1 - \rho^+_r)} = \frac{\lambda^+_r - \lambda^+_r (1 - \rho)}{\lambda(1 - \rho^+_r)} \]

\[
= \frac{(1 - \rho)(\lambda^+_r - \lambda^+_r (1 - \rho^+_r))}{\lambda(1 - \rho^+_r)}
\]

\[
= \frac{1 - \rho}{\rho} \left( \frac{1 - \rho^+_r}{1 - \rho^+_r} - \frac{1 - \rho^+_r}{1 - \rho^+_r} \right).
\]

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1062.

Acknowledgments

The authors thank the editors and referees for their helpful comments. The authors are grateful to Joseph Milner for his useful feedback on an earlier version of this paper. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

Abouee-Mehrizi, Balcıoğlu, and Baron: A Centralized Multiclass M/G/1 Queue


Hossein Abouee-Mehrizi is an assistant professor of management sciences at the University of Waterloo. He is broadly interested in advanced stochastic processes and queueing analysis.

Barış Balcıoğlu is an associate professor of industrial engineering at the Faculty of Engineering and Natural Sciences, Sabancı University. His research and teaching activities are in the areas of queueing theory, discrete-event simulation, performance analysis of call centers and production/inventory systems, and reliability and maintenance engineering.

Opher Baron is an associate professor of operations management at the Rotman School of Management, the University of Toronto. He has a Ph.D. in operations management from the Massachusetts Institute of Technology Sloan School of Management along with an M.B.A. and a B.Sc. in industrial engineering and management from the Technion–Israel Institute of Technology. His research interests include applied probability and its application to facility location, service operations, inventory planning, and revenue management.